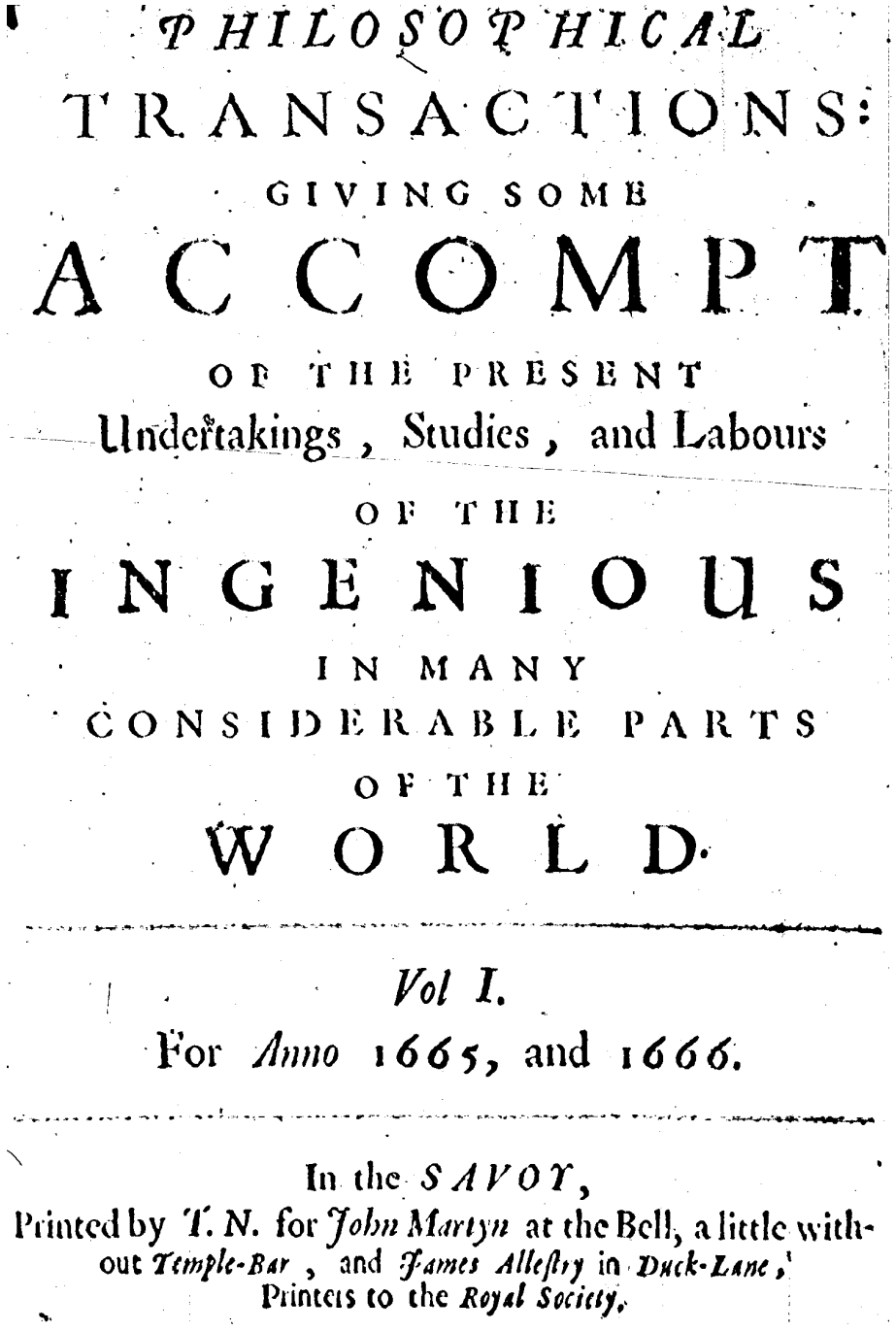
National Snow and Ice Data Center
University of Colorado at Boulder

Data Citation

Mark Parsons

ESIP Winter Meeting
4 January 2011



PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.

Vol I.
For Anno 1665, and 1666.

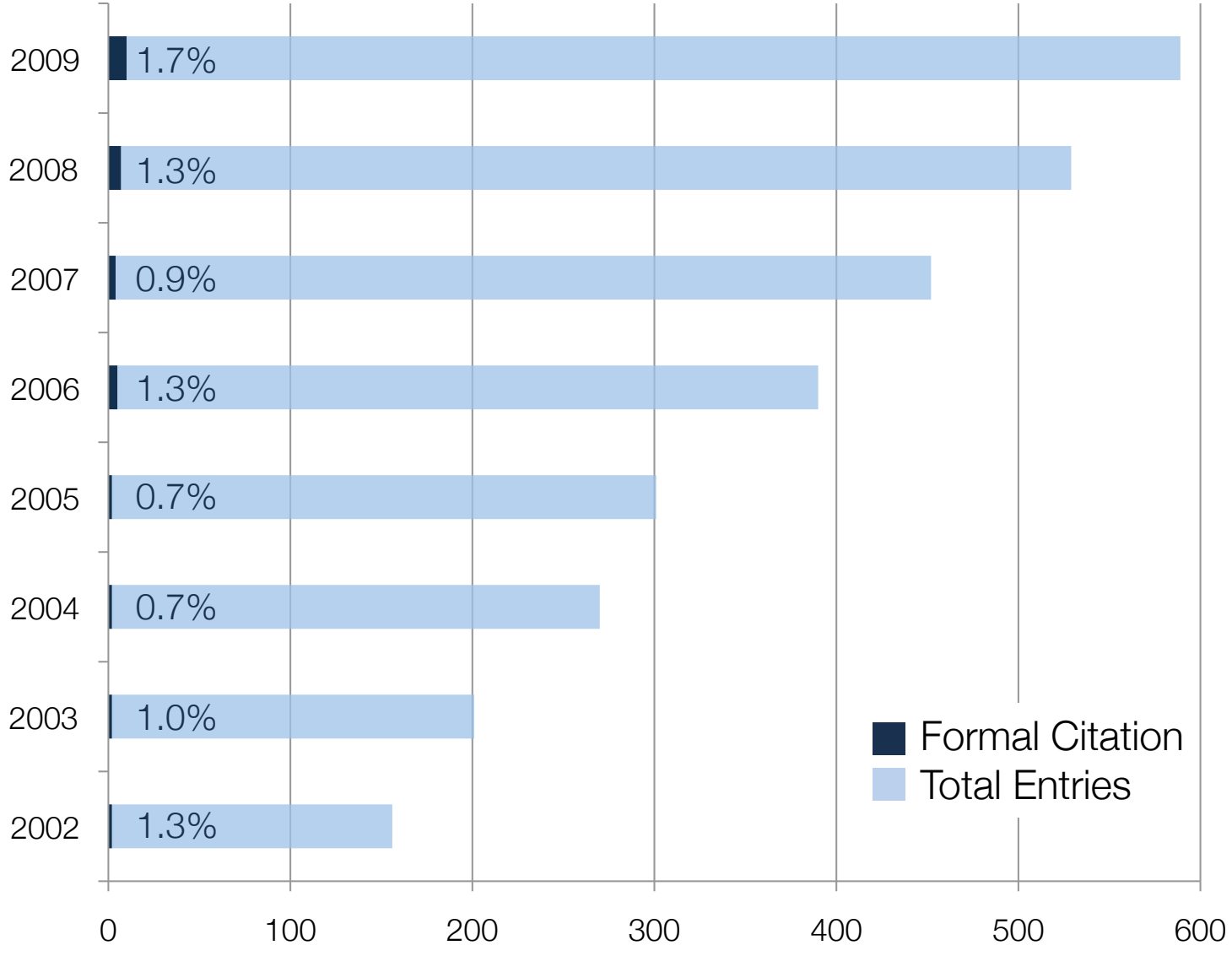
In the SAVOY,
Printed by T. N. for John Martyn at the Bell, a little without Temple-Bar, and James Allestry in Duck-Lane,
Printers to the Royal Society.

Purpose of Data Citation

- Credit for data creators
- Track impact of data set
- Accountability
- Aid reproducibility through direct, unambiguous connection to the precise data used.

How data citation is currently done

- Not mentioned, just used, e.g., in tables or figures
- URL in text (with variable degrees of specificity)
- Reference to name or source of data in text
- Citation of related paper (e.g. CRU Temp. records recommend citing two old journal articles which do not contain the actual data of full description of methods)
- Citation of actual data set typically using recommended citation given by data center
- Citation of data set including a persistent identifier/locator, typically a DOI



"MODIS Snow Cover Data" in Google Scholar

The National Snow and Ice Data Center distributes a variety of different snow cover products derived from the Moderate Resolution Imaging Spectroradiometer (MODIS). The results of a quick analysis of how many scientific papers mention use of "MODIS Snow Cover Data" (according to Google Scholar) and how often the data sets themselves are formally cited show a huge disparity, illustrating the infrequency of proper data citation in practice. Moreover, the lack of data citation standards introduces the possibility that informal references to data do not point to the exact data set actually used.

Data Citation Guidelines

- International Polar Year — http://ipydis.org/data/citations.html
- DataCite—a consortium of libraries and related organizations working to define a citation approach around DOIs. Schema is out for review.
- New CODATA Task Group in collaboration with ICSTI
- DataVerse Network Project—a standard from the social science community using a Handle locator and “Universal Numerical Fingerprint” as a unique identifier.
- Most NASA DAACs and other data centers but with great variation in approach
- USGS generally requests acknowledgement, but maps are cited, and a more formal approach was proposed.
- NOAA National Data Centers simply request acknowledgement
- Overall approaches range from specific data citation, to general acknowledgement, to recommending citing a journal article or even a presentation.

“Data Citation in the Wild”

Valerie Enriquez, Sarah Walker Judson, Nicholas M. Weber, Suzie Allard, Robert B. Cook, Heather A. Piwowar, Robert J. Sandusky, Todd J. Vision, Bruce Wilson

“We found that few policies recommend robust data citation practices: in our preliminary evaluation, only one-third of repositories (n=26), 6% of journals (n=307), and 1 of 53 funders suggested a best practice for data citation. We manually reviewed 500 papers published between 2000 and 2010 across six journals; of the 198 papers that reused datasets, only 14% reported a unique dataset identifier in their dataset attribution, and a partially-overlapping 12% mentioned the author name and repository name. Few citations to datasets themselves were made in the article references section.”

http://openwetware.org/wiki/DataONE:Notebook/Summer_2010

Need two strategies

- Publishers and educators need to provide guidelines and requirements on how to cite data, esp. historical data.
- Data Centers need to provide consistent, precise citation recommendations for their data.

An Example Citation

Cline, D., R. Armstrong, R. Davis, K. Elder, and G. Liston. 2002, Updated July 2004. *CLPX-Ground: ISA snow pit measurements.* Edited by M. Parsons and M. J. Brodzik. Boulder, CO: National Snow and Ice Data Center. Data set accessed 2008-05-14 at http://nsidc.org/data/nsidc-0176.html.

An Example Citation

Cline, D., R. Armstrong, R. Davis, K. Elder, and G. Liston. 2002, Updated July 2004. *CLPX-Ground: ISA snow pit measurements.* Edited by M. Parsons and M. J. Brodzik. Boulder, CO: National Snow and Ice Data Center. Data set accessed 2008-05-14 at http://nsidc.org/data/nsidc-0176.html.

An Example Citation

Cline, D., R. Armstrong, R. Davis, K. Elder, and G. Liston. 2002, Updated July 2004. *CLPX-Ground: ISA snow pit measurements.* Edited by M. Parsons and M. J. Brodzik. Boulder, CO: National Snow and Ice Data Center. Data set accessed 2008-05-14 at http://nsidc.org/data/nsidc-0176.html.

An Example Citation

Cline, D., R. Armstrong, R. Davis, K. Elder, and G. Liston. 2002, Updated July 2004. *CLPX-Ground: ISA snow pit measurements.* Edited by M. Parsons and M. J. Brodzik. Boulder, CO: National Snow and Ice Data Center. Data set accessed 2008-05-14 at http://nsidc.org/data/nsidc-0176.html.

An Example Citation

Cline, D., R. Armstrong, R. Davis, K. Elder, and G. Liston. 2002, Updated July 2004. *CLPX-Ground: ISA snow pit measurements.* Edited by M. Parsons and M. J. Brodzik. Boulder, CO: National Snow and Ice Data Center. Data set accessed 2008-05-14 at http://nsidc.org/data/nsidc-0176.html.

An Example Citation

Cline, D., R. Armstrong, R. Davis, K. Elder, and G. Liston. 2002, Updated July 2004. *CLPX-Ground: ISA snow pit measurements.* Edited by M. Parsons and M. J. Brodzik. Boulder, CO: National Snow and Ice Data Center. Data set accessed 2008-05-14 at http://nsidc.org/data/nsidc-0176.html.

An Example Citation

Cline, D., R. Armstrong, R. Davis, K. Elder, and G. Liston. 2002, Updated July 2004. *CLPX-Ground: ISA snow pit measurements.* Edited by M. Parsons and M. J. Brodzik. Boulder, CO: National Snow and Ice Data Center. Data set accessed 2008-05-14 at http://nsidc.org/data/nsidc-0176.html.

A note on versions and time series

Hall, Dorothy K., George A. Riggs, and Vincent V. Salomonson. 2006, updated daily. *MODIS/Terra snow cover Extent 5-Min L2 swath 1km* V005, Oct. 2007–Apr. 2008. Boulder, Colorado USA: National Snow and Ice Data Center. Data set accessed 2008-05-14 at http://nsidc.org/data/myd29v5.html.

Hall, D. K., G. A. Riggs, and V. V. Salomonson. 2000, updated daily. *MODIS/Terra snow cover 5-Min L2 swath 500m V004, Oct. 2007–Apr. 2008.* Boulder, Colorado USA: National Snow and Ice Data Center. Data set accessed 2008-05-14 at http://nsidc.org/data/myd29v5.html.

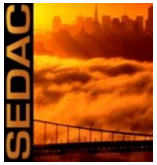
An Example Citation

König-Langlo, Gert and Hatwig Gernandt. 2006. *Compilation of radiosonde data from the Antarctic Georg-Forster station of the German Democratic Republic from 1985 to 1992*. Bremerhaven, Germany: Alfred Wegener Institute for Polar and Marine Research Data set accessed 2008-05-22. doi:10.1594/PANGAEA.547983

An Example Citation

Gary King; Langche Zeng, 2006, "Replication Data Set for 'When Can History be Our Guide? The Pitfalls of Counterfactual Inference'" hdl:1902.1/DXRXCFAWPK UNF:3:DaYIT6QSX9r0D50ye+tXpA== Murray Research Archive [distributor]

Some measurement issues (Chen and Downs 2010)



- Data are not used in isolation – often different data are combined, used with models and other analytic techniques
- Impacts may be indirect, i.e., resulting from development of information, papers, tools, etc. that relied on derived data or products
- Impacts may be delayed, i.e., months or years for a peer-reviewed publication to be released, or a decision to be made and implemented
- Impacts may be unexpected, e.g., a new scientific discovery or a novel application of data collected for a different purpose
- Impacts may be hard to compare, e.g., in scientific, economic, or ethical terms

But it is still important to try:

- **Need to justify investment in data acquisition, maintenance, distribution and long-term stewardship!**
- **Need to help community become more effective and efficient in data management and use!**

Chen, R. S. and Downs, R. R. (2010). Evaluating the Use and Impacts of Scientific Data. National Federation of Advanced Information Services (NFAIS) Workshop, Assessing the Usage and Value of Scholarly and Scientific Output: An Overview of Traditional and Emerging Approaches. Philadelphia, PA, November 10, 2010. http://info.nfais.org/info/ChenDownsNov10.pdf

Possible citation metrics

- Qualitative
  - Examples of data use and impacts in key papers, discoveries, decisions
  - Assessment of broader impacts such as influence of data on attitudes and thinking (e.g., Apollo 8 image!)
- Quantitative
  - Counts of papers that cite data in peer-reviewed journals
  - Weighted indicators of data citations (e.g., type/quality of citation, impact of journal)
- Quantitative and Qualitative:
  - Number of data citations in top peer-reviewed scientific journals and “key” reports by decision-makers
  - Data usage in other peer-reviewed journals, textbooks, reports, magazines, documentary films, online tools, maps, blogs, twitters, etc.

Tracking citation

"Tracking Dataset Citations Using Common Citation Tracking Tools Doesn't Work"

—Heather Pinowar, DataONE

- Traditional fields such as author and date too imprecise
- Web of Science, Scopus, and other tools don't handle identifiers

Accountability

- “A new standard of accountability” -Tom Moritz
- Data “publication” needs to be tied to promotion, tenure, etc.
- Implies peer review— See AGU Position Statement on Data
- What is peer-review?
  - An assertion of accuracy or validity?
  - An audit of complete documentation and sound practice?
  - Related to but different than QA.
  - How does it overlap with curation and stewardship?
- *Earth System Science Data* one approach, but not universally applicable.
- Open or informal review or usage comments within the metadata
- Versioning and transparency are essential

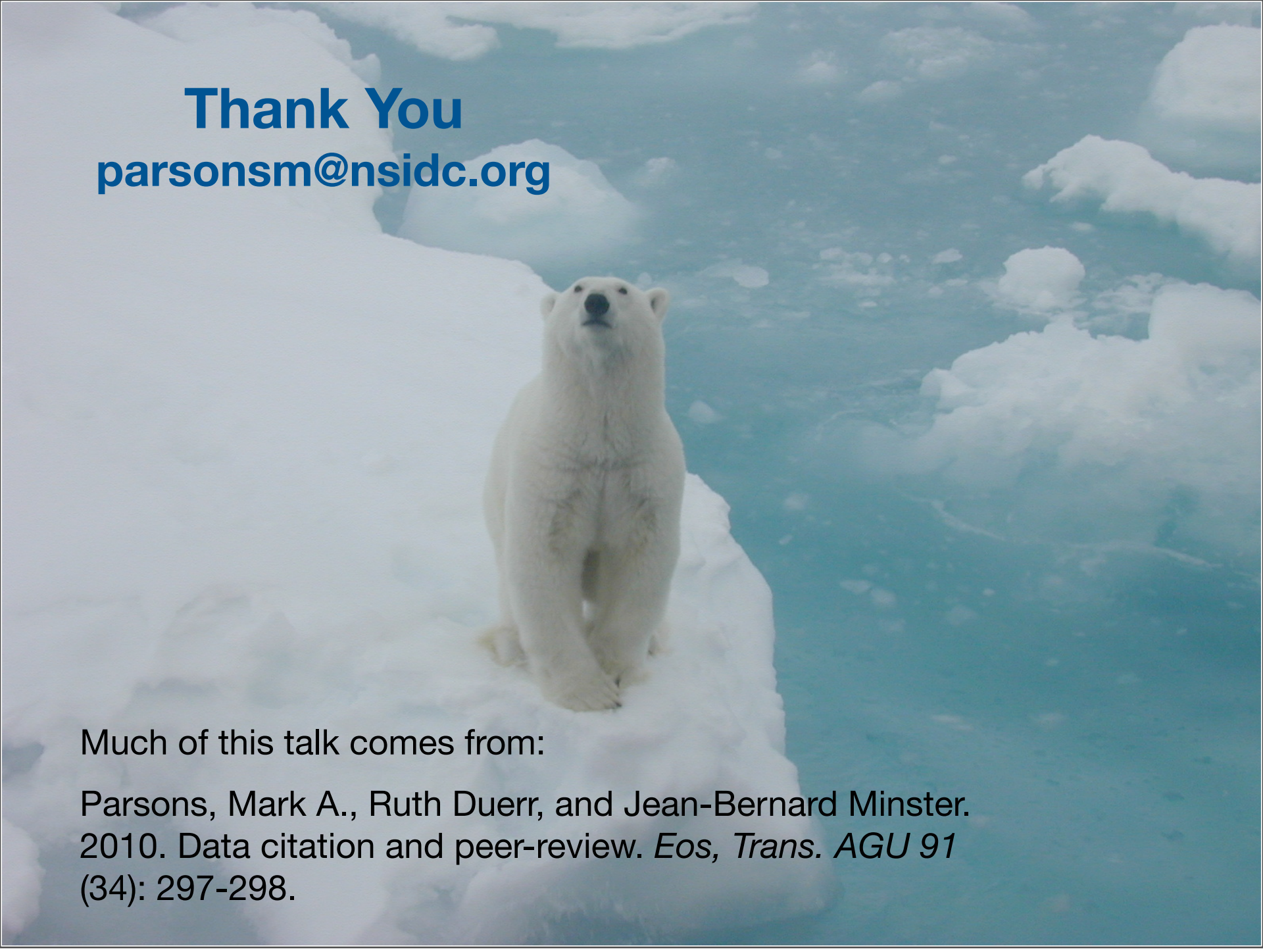
Identifiers

- DOIs increasingly used.
  - Not perfect but well understood by publishers
  - DataCite working with Thomson Reuters
  - What is the citable unit?
  - Versioning
  - “Retired” data
- To be able to tell that two files contain the same data even if the formats are different. That is, how do we ensure that two files are “scientifically identical”?
  - Universal Numeric Fingerprint?

What is the citable unit with a DOI? A file? A collection of files? How many? Further, it is important to note that data products can be purged from an archive; such deleted information still needs to be able to be referenced. Even if the products themselves are not preserved, the raw data must be preserved along with detailed documentation describing how the product was created, and that documentation must be citable.

Suggestions

- Collaborate with CODATA Task Group, DataCite, and GEOSS
- Use IPY Guidelines for now, especially for existing works.
- Apply DOIs (or ARKs or PURLs) to collections, consider versioning and data stability.
- Have systems create, attach, and maintain UUIDs with files.
- Explore the use of the Universal Numeric Fingerprint
- Apply machine readable “badges” (e.g. CC0)
- Encourage publishers to require data citation—a new norm of science.



Thank You

parsonsm@nsidc.org

Much of this talk comes from:

Parsons, Mark A., Ruth Duerr, and Jean-Bernard Minster. 2010. Data citation and peer-review. *Eos, Trans. AGU 91* (34): 297-298.